Data 583 Life Expectancy (WHO)

Justin Chan, Kenny Tong, Viji Rajagopalan

7 Mar, 2023

EDA

Original Dataset Summary & Initial Data Screening

Purpose : Let's take a snapshot of the original dataset and have a rough idea of its record

```
le <- read.csv("dataset/LifeExpectancy.csv")
summary(le)
```

```
##    Country               Year         Status          Life.expectancy
##  Length:2938        Min.   :2000   Length:2938        Min.   :36.30
##  Class :character   1st Qu.:2004   Class :character   1st Qu.:63.10
##  Mode  :character   Median :2008   Mode  :character   Median :72.10
##                     Mean   :2008                      Mean   :69.22
##                     3rd Qu.:2012                      3rd Qu.:75.70
##                     Max.   :2015                      Max.   :89.00
##                                                       NA's   :10
##  Adult.Mortality infant.deaths       Alcohol        percentage.expenditure
##  Min.   :  1.0   Min.   :   0.0   Min.   : 0.0100   Min.   :    0.000
##  1st Qu.: 74.0   1st Qu.:   0.0   1st Qu.: 0.8775   1st Qu.:    4.685
##  Median :144.0   Median :   3.0   Median : 3.7550   Median :   64.913
##  Mean   :164.8   Mean   :  30.3   Mean   : 4.6029   Mean   :  738.251
##  3rd Qu.:228.0   3rd Qu.:  22.0   3rd Qu.: 7.7025   3rd Qu.:  441.534
##  Max.   :723.0   Max.   :1800.0   Max.   :17.8700   Max.   :19479.912
##  NA's   :10                       NA's   :194
##   Hepatitis.B       Measles               BMI        under.five.deaths
##  Min.   : 1.00   Min.   :     0.0   Min.   : 1.00   Min.   :   0.00
##  1st Qu.:77.00   1st Qu.:     0.0   1st Qu.:19.30   1st Qu.:   0.00
##  Median :92.00   Median :    17.0   Median :43.50   Median :   4.00
##  Mean   :80.94   Mean   :  2419.6   Mean   :38.32   Mean   :  42.04
##  3rd Qu.:97.00   3rd Qu.:   360.2   3rd Qu.:56.20   3rd Qu.:  28.00
##  Max.   :99.00   Max.   :212183.0   Max.   :87.30   Max.   :2500.00
##  NA's   :553                        NA's   :34
##      Polio       Total.expenditure   Diphtheria       HIV.AIDS
##  Min.   : 3.00   Min.   : 0.370    Min.   : 2.00   Min.   : 0.100
##  1st Qu.:78.00   1st Qu.: 4.260    1st Qu.:78.00   1st Qu.: 0.100
##  Median :93.00   Median : 5.755    Median :93.00   Median : 0.100
##  Mean   :82.55   Mean   : 5.938    Mean   :82.32   Mean   : 1.742
##  3rd Qu.:97.00   3rd Qu.: 7.492    3rd Qu.:97.00   3rd Qu.: 0.800
##  Max.   :99.00   Max.   :17.600    Max.   :99.00   Max.   :50.600
##  NA's   :19      NA's   :226       NA's   :19
##       GDP              Population        thinness..1.19.years
##  Min.   :     1.68   Min.   :3.400e+01   Min.   : 0.10
```

```
##  1st Qu.:   463.94   1st Qu.:1.958e+05   1st Qu.: 1.60
##  Median :  1766.95   Median :1.387e+06   Median : 3.30
##  Mean   :  7483.16   Mean   :1.275e+07   Mean   : 4.84
##  3rd Qu.:  5910.81   3rd Qu.:7.420e+06   3rd Qu.: 7.20
##  Max.   :119172.74   Max.   :1.294e+09   Max.   :27.70
##  NA's   :448         NA's   :652         NA's   :34
##  thinness.5.9.years Income.composition.of.resources   Schooling
##  Min.   : 0.10      Min.   :0.0000                  Min.   : 0.00
##  1st Qu.: 1.50      1st Qu.:0.4930                  1st Qu.:10.10
##  Median : 3.30      Median :0.6770                  Median :12.30
##  Mean   : 4.87      Mean   :0.6276                  Mean   :11.99
##  3rd Qu.: 7.20      3rd Qu.:0.7790                  3rd Qu.:14.30
##  Max.   :28.60      Max.   :0.9480                  Max.   :20.70
##  NA's   :34         NA's   :167                     NA's   :163
```

Let's look at the dataset dimension first

```
dim(le)
```

```
## [1] 2938   22
```

Then, have a quick overall screening of the dataset

```
head(le,5)
```

```
##       Country Year     Status Life.expectancy Adult.Mortality infant.deaths
## 1 Afghanistan 2015 Developing            65.0             263            62
## 2 Afghanistan 2014 Developing            59.9             271            64
## 3 Afghanistan 2013 Developing            59.9             268            66
## 4 Afghanistan 2012 Developing            59.5             272            69
## 5 Afghanistan 2011 Developing            59.2             275            71
##   Alcohol percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths
## 1    0.01              71.279624          65    1154 19.1                83
## 2    0.01              73.523582          62     492 18.6                86
## 3    0.01              73.219243          64     430 18.1                89
## 4    0.01              78.184215          67    2787 17.6                93
## 5    0.01               7.097109          68    3013 17.2                97
##   Polio Total.expenditure Diphtheria HIV.AIDS       GDP Population
## 1     6              8.16         65      0.1 584.25921   33736494
## 2    58              8.18         62      0.1 612.69651     327582
## 3    62              8.13         64      0.1 631.74498   31731688
## 4    67              8.52         67      0.1 669.95900    3696958
## 5    68              7.87         68      0.1  63.53723    2978599
##   thinness..1.19.years thinness.5.9.years Income.composition.of.resources
## 1                 17.2                17.3                           0.479
## 2                 17.5                17.5                           0.476
## 3                 17.7                17.7                           0.470
## 4                 17.9                18.0                           0.463
## 5                 18.2                18.2                           0.454
##   Schooling
## 1      10.1
## 2      10.0
## 3       9.9
## 4       9.8
## 5       9.5
```

Here is another view :

```
str(le)
```

```
## 'data.frame':    2938 obs. of  22 variables:
##  $ Country                        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
##  $ Status                         : chr  "Developing" "Developing" "Developing" "Developing" ...
##  $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
##  $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
##  $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
##  $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
##  $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
##  $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
##  $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
##  $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
##  $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
##  $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
##  $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
##  $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
##  $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
##  $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
##  $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
##  $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405 ...
##  $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

From the above broad view, the following Conclusion/Key Findings are reached :

- The records range is from Year 2000 to 2015
- Columns with NA : Life Expectancy, Adult Mortality, Alcohol, Hep B, BMI, Polio, Total exp, Dip, GDP, Population, thinness..1.19, thinness.5.9, Income.composition.of.resources, Schooling
- 'Status' Column is of the "character" data type, with values "Developing" and "Developed". We will introduce a new column 'Status.val' to be the factor value of 'Status' for better analysis..
- 'Percentage Expenditure' has a mean value of 738.2512955 and max. value of 1.9479912×10^4. Spending on health is more than the GDP per capita? Look into the column definition : Expenditure on health as a percentage of Gross Domestic Product per capita(%). The data of such magnitude simply does not quite make sense. Cross check with other references (e.g. the World Bank https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS). OK, let's conclude that we have hesitation about the reliability/interpretation of the value of this column, and probably would drop and skip this column for the rest of this analysis.
- 'Population' and 'GDP' have a relatively large scale, compared with all other columns. So, we may need to scale these two columns.

Now, let's do some data wrangling based on the above conclusions :

```
# Create a new column Status.val to represent the Status column with number
le$Status.val <- ifelse(le$Status == "Developed",1,0)

# Create a new column as the scaled version of the GDP & Population,
le$GDP_scaled = scale(le$GDP)
le$Population_scaled = scale(le$Population)

# Remove the unreliable column
le <- subset(le,select=-c(percentage.expenditure))
```

Null Value Analysis and Handling

Purpose : Investigate the and determine how to handle the null value in the data set

There are 2938 no. of rows in the dataset. Let's set the threshold of 20% as the max. proportion of null column to be allowed in a data column. That means, columns with na over 20% will be dropped. The threshold is then 587.6. So, the following 'Population' column will be dropped.

```
#head(le)
le <- subset(le,select=-c(Population))
# also Population_Scaled
le <- subset(le,select=-c(Population_scaled))
```

For the other values, we will set the na to the respective column mean for the subsequent analysis.

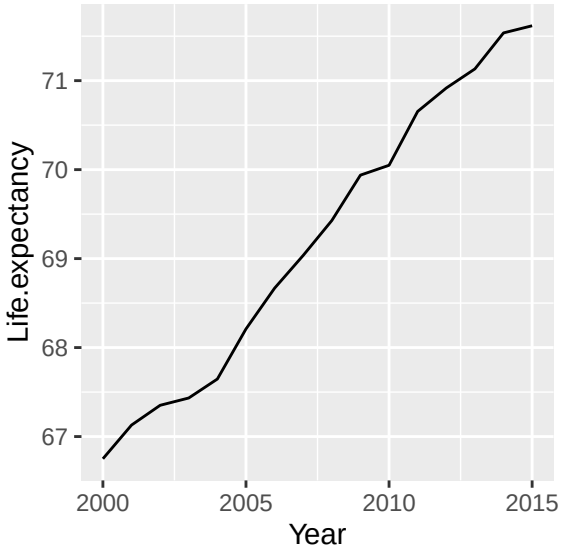
```
for(i in 1:ncol(le)) {                                    # Replace NA in all columns
  le[ , i][is.na(le[ , i])] <- mean(le[ , i], na.rm = TRUE)
}
```

Now, all na have been handled! Let's continue our analysis.

Overall General Life Expectancy Trend

Purpose : Do some visualisation to explore and identify the general data pattern, trends and clusters, etc

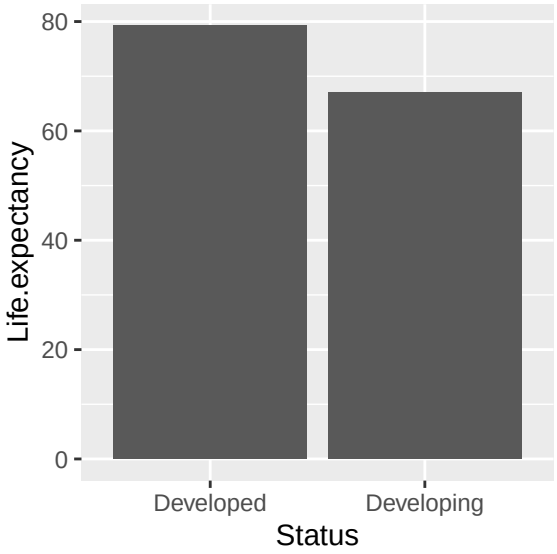
```
library(ggplot2)
library(tidyverse)

le %>%
  group_by(Year) %>%
  summarise(Life.expectancy = mean(Life.expectancy)) %>%
  ggplot(aes(x=Year,
             y=Life.expectancy)) +
  geom_line()
```



Findings :

- The general life expectancy has been steadily increasing duration the year
- Average Life expectancy increase from about 67 to 71.5 in 15 years.

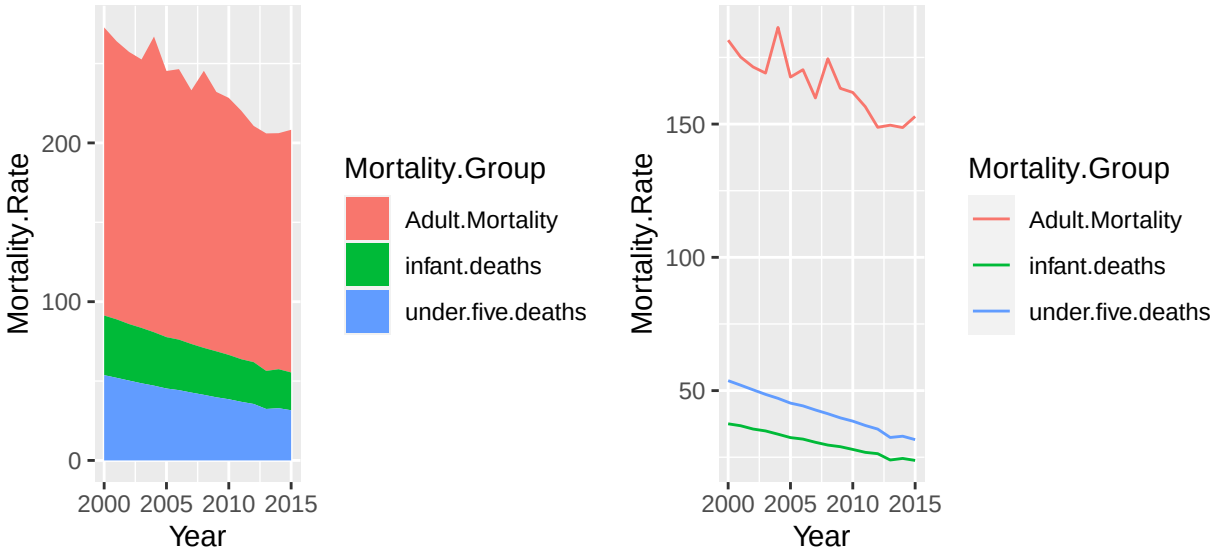
```
le %>%
  group_by(Status) %>%
  summarise(Life.expectancy = mean(Life.expectancy)) %>%
  ggplot(aes(x=Status,
             y=Life.expectancy)) +
  geom_bar(stat = "identity")
```



Finding :

- Life expectancy of Developed countries are significantly higher than that of Developing countries.

```
le.pivot <- pivot_longer(le,c(Adult.Mortality,under.five.deaths,infant.deaths),names_to='Mortality.Group',
require(gridExtra)

le.pivot.area <- le.pivot %>%
  group_by(Year,Mortality.Group) %>%
  summarise(Mortality.Rate = mean(Mortality.Rate)) %>%
  ggplot(aes(x=Year,
             y=Mortality.Rate,
             fill=Mortality.Group)) +
  geom_area(position="stack",stat="identity")

le.pivot.line <- le.pivot %>%
  group_by(Year,Mortality.Group) %>%
  summarise(Mortality.Rate = mean(Mortality.Rate)) %>%
  ggplot(aes(x=Year,
             y=Mortality.Rate,
             color=Mortality.Group)) +
  geom_line()

grid.arrange(le.pivot.area, le.pivot.line, ncol=2)
```



Findings :

- The mortality rate of all three age groups are generally decreasing as a whole
- The mortality rate of the adult group, however, have fluctuation within the period